We will define one **person** of compute as 20 PFLOPS (64 A100s, or a single dense 42U A100 rack). We are in the era of the 1 rack person, consuming about 30kW to provide those 20 PFLOPS.

LLaMA was trained on a cluster of 2048 A100s, with ~312 TFLOPS each. 2048 is currently the most A100s that can work together on a model due to the switch topology.

The cluster has 639 PFLOPS, or 32 people of compute. Large LLaMA used ~1M GPU hours to train. Meaning it used the cluster for 500 hours (3 weeks). 32 people for 3 weeks is about 2 person-years of work. GPT-4 was about 100 person-years

As long as these things are around human scale (for the next 10 years), I think these units make sense.

One way to think about scaling computers is how many Moore's laws it gets you over a desktop. A desktop today is ~50 TFLOPS, something like a mouse of compute, one 400th of a person, or nine more Moores. (2038 remains my lifelong estimate for what most people consider the Singularity. I like that it's the Unix timestamp rollover)

Some NVIDIA numbers:

- 1080 (2016) = 11.3 TFLOPS
- 2080 (2018) = 14.2 TFLOPS
- 3090 (2020) = 35.6 TFLOPS
- 4090 (2022) = 82.6 TFLOPS

The 4090 cheated a bit by using tons of power, but overall we're on track for a doubling every two years.

Today, you can buy two years by doubling your budget. Facebook is log2(2048) = 11 Moores, or ~16 years ahead (you can afford an 8 GPU box, right?).

Google is claiming they have a 9 exaflop (450 person) computer. They bought 8 more years, that's a computer from 24 years in the future.

On a tight budget, a person of compute costs about $250k today. That's a $115M computer. The most expensive thing humanity has built is the ISS at $100B. If we built a computer at that scale, it would be 400,000 people.

That's one Tampa of compute, which is the most we could hope to build today. A single Tampa.

One *Humanity* is 20,000 Tampas.

- In 24 years, we can build a Humanity for the cost of the ISS.
- In 44 years, future Google will have a Humanity.
- In 54 years, a normal sized cluster will be a Humanity.
- And in 66 years, you'll have a Humanity under your desk.

Ugh this needs to happen sooner.